# Social risk factors and cardiovascular risk in obstructive sleep apnea: a systematic assessment of clinical predictors in community health centers[a]

Diego R. Mazzotti[1]; Ryan Urbanowicz[2]; Marta Jankowska[3]

*[1]Division of Medical Informatics, Division of Pulmonary Critical Care and Sleep Medicine, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, United States Email: droblesmazzotti@kumc.edu; [2]Department of Computational Biomedicine, Cedars Sinai Medical Center, Los Angeles, CA; Department of Biostatistics Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, United States, Email: Ryan.Urbanowicz@cshs.org; [3]Population Sciences, Beckman Research Institute, City of Hope, Duarte, CA, United States Email: mjankowska@coh.org*

We leveraged electronic health record (EHR) data from the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN) to identify social risk factor clusters, assess their association with obstructive sleep apnea (OSA), and determine relevant clinical predictors of cardiovascular (CV) outcomes among those experiencing OSA. Geographically informed social indicators were used to define social risk factor clusters via latent class analysis. EHR-wide diagnoses were used as predictors of 5-year incidence of major adverse CV events (MACE) using STREAMLINE, an end-to-end rigorous and interpretable automated machine learning pipeline. Analyses among over 1.4 million individuals revealed three major social risk factor clusters: lowest (35.7%), average (43.6%) and highest (22.7%) social burden. In adjusted analyses, those experiencing highest social burden were less likely to have received a diagnosis of OSA when compared to those experiencing lowest social burden (OR [95%CI]=0.85[0.82-0.88]). Among those with OSA and free of prior CV diseases (N=4,405), performance of predicting incident MACE reached a ROC-AUC of 0.70 [0.03] overall but varied when assessed within each social risk factor cluster. Feature importance also revealed that different clinical factors might explain predictions among each cluster. Results suggest relevant health disparities in the diagnosis of OSA and across clinical predictors of CV diseases among those with OSA, across social risk factor clusters, indicating that tailored interventions geared toward minimizing these disparities are warranted.

*Keywords:* Health disparities; Social risk factors; sleep disorders; cardiovascular risk; electronic health records.

[a] This research was, in part, funded by the National Institutes of Health (NIH) Agreement NO. 1OT2OD032581-01. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NIH. The research reported in this work was powered by PCORnet®. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute® (PCORI®) and conducted with the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN). ADVANCE is a Clinical Research Network in PCORnet® led by OCHIN in partnership with Health Choice Network, Fenway Health, University of Washington, and Oregon Health & Science University. ADVANCE's participation in PCORnet® is funded through the PCORI Award RI-OCHIN-01-MC.

## 1. Introduction

Sleep problems disproportionally affect populations experiencing health disparities[1]. Racial, ethnic and socioeconomically disadvantaged minorities are more likely to experience insufficient sleep[2-8], sleep disorders[9], and negative cardiovascular (CV) outcomes[10,11]. Yet, many of these conditions go unnoticed in these populations, largely due to lack of healthcare access focused on diagnosing and treating sleep disorders. Consequently, pathways linking health disparities to sleep disturbances and CV outcomes are largely underexplored, particularly among underrepresented populations.

Obstructive sleep apnea (OSA) is a heterogeneous sleep disordered breathing condition and one of the most prevalent sleep disorders, affecting approximately 1 billion adults worldwide[12]. Epidemiological and experimental evidence supports a major role of OSA towards increasing CV risk[13-16]. However, prior studies were mostly focused on population or community-based cohorts that generally underrepresented important groups known to be at greater risk of experiencing health disparities. The identification of clinical predictors of major adverse CV events (MACE) in these populations is a necessary step towards design tailored and equitable sleep-promoting interventions towards improved CV health.

Efforts supporting the integration and availability of electronic health record (EHR) data linked with relevant social risk information is essential to better characterize the effects of health disparities. Towards that goal, initiatives such as the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN) led by the OCHIN network of community health organizations enable such studies[17], with a great potential to inform public health. As such, the current study leveraged data from the ADVANCE CRN and demonstrated an approach to dissect the heterogeneity of geographically informed social risk factors by applying clustering techniques and identifying social risk factor clusters. This data-driven approach supports the identification of population subgroups experiencing similar levels of social exposures and can offer an exploratory perspective on the impact of socio-environmental burden on health. We further assessed the association between social risk factors clusters and evidence of OSA diagnosis. Next, by employing a robust, end-to-end, and interpretable automated machine learning (ML) pipeline, we assessed clinical predictors of 5-year incidence of new onset MACE among individuals with OSA belonging to different clusters. We hypothesized that 1) individuals experiencing higher social burden were less likely to have received a diagnosis of OSA; and 2) clinical predictors of incident MACE varied across social risk clusters, likely reflecting different pathways towards CV risk depending on socio-environmental exposure.

## 2. Methods

### 2.1. *Study Design and Population*

This is a retrospective clinical cohort study of patients at risk for sleep disorders that were part of the ADVANCE CRN with available geographically informed social risk factor data ascertained between 2012 and 2021. Data was sourced from the OCHIN Epic EHR system. Data is representative of outpatient community-based health care organizations delivering high-quality primary care services for communities impacted by health disparities in the U.S. Clinical institutions

include Federally Qualified Health Centers or other federally supported community health centers. The ADVANCE CRN is part of PCORnet®, the National Patient-Centered Clinical Research Network, thus data is organized according to the PCORnet® Common Data Model. Access was requested and facilitated by the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program. The study has been approved by Institutional Review Boards from the University of Kansas Medical Center and Harvard Medical School with non-human subjects determination, as only de-identified data was made available.

Out of a dataset of over 3.2 million adults (age ≥18 years), we identified a cohort with at least one year of interactions with community health centers, a minimum of 3 encounters, and non-missing geographically informed social risk factors. Among those patients, we further created a subset of those with evidence of OSA and at least 5 years of interaction with the community health centers and without prior evidence of CV diseases to determine clinical predictors of incident MACE.

### 2.2. *Geographically informed social risk factors*

Geographically linked neighborhood-level indicators at census tract and/or ZCTA levels[18] were made available through OCHIN as part of the ADVANCE CRN data warehouse. Linkage was performed by matching participant's address ZIP code with publicly available data sources from the U.S. Census Bureau and American Community Survey, and used to impute the following area-level social indicators: income inequality coefficient, or Gini coefficient, a measure ranging from 0 (perfectly equal geographical region where all income is equally shared) and 1 (perfectly unequal society where all income is earned by 1 individual)[19,20]; median household income (in U.S. dollars); percent of adults age >25 years who graduated from college; percent of total population in poverty (<100% federal poverty level [FPL]); and rate of unemployment among population age ≥16 years. These indicators were categorized into quartiles prior to downstream analyses.

### 2.3. *Computable phenotypes for OSA*

A validated EHR algorithm was used to identify individuals with evidence of OSA, as described by Keenan et al. 2020[21]. Individuals with 2 or more International Classification of Diseases (ICD)-9 or 10 codes for OSA at different dates were classified as having OSA (ICD-9: 327.20, 327.23, 327.29, 780.51, 780.53, 780.57; ICD-10: G47.30, G47.33, G47.39). This algorithm presented optimal predictive performance across six health systems in the U.S., with overall positive predictive value (95% CI) of 97.1% (95.6, 98.2) and negative predictive value of 95.5% (93.5, 97.0)[21]. Individuals not meeting these criteria were defined as not having evidence of OSA diagnosis.

### 2.4. *Phecode mapping*

The phecode framework[22] is a high-throughput EHR phenotyping method with the goal of representing a wide range of clinical phenotypes. Structured as an ontology-based classification system, phecodes combine groups of ICD codes into clinically relevant groups, thus minimizing the dimensionality of clinical diagnosis. In this study, we focused on phecodes observed in at least 1,000

participants in our final cohort, resulting in a total of 932 phecodes included as predictors in our ML analyses. Phecode maps can be queried elsewhere (https://www.phewascatalog.org/phecodes).

### 2.5. *Study outcomes*

We report the results of two analyses. Our primary analysis consisted of investigating the association between social risk factors clusters and evidence of OSA. Thus, our outcome was prevalence of OSA. Secondly, we assessed clinical predictors of 5-year incidence of MACE, defined as a composite of myocardial infarction, coronary artery disease, cerebrovascular disease, heart failure or stroke, using validated computable phenotypes as previously described[23-29]. A list of ICD and Current Procedural Terminology codes used to define these conditions are available elsewhere (https://raw.githubusercontent.com/RWD2E/phecdm/main/res/valueset_curated/vs-osa-comorb.json).

### 2.6. *Statistical analyses*

All analyses were conducted within the AIM-AHEAD Service Workbench cloud infrastructure. Initial cohort characterization was performed through a data request with OCHIN. A database schema was created in Microsoft SQL Server and access was provided to the author. A series of tables resulting from this database schema were generated to capture the following data domains: patient demographics, social risk factors, diagnosis, and procedures. Queries used to create analysis-ready can be found elsewhere (https://github.com/mazzottidr/AIMAHEAD_Fellowship_Mazzotti).

First, we determined univariate associations between OSA and sociodemographic characteristics (sex, race, ethnicity, gender identity, current FPL, marital status, homeless status, and sexual orientation), as well as between OSA and quartiles of geographically informed social risk factors (Gini coefficient, median household income, percent of college graduates; percent of total population in poverty; and rate of unemployment) using chi-squared tests or t tests. Next, we used latent class analysis (LCA) to identify clusters of social risk factors using quartiles of the geographically informed social risk factors listed above. Due to the large computational requirements of performing LCA on large datasets, we assessed the optimal number of clusters by sub-setting the data into 10 random subsamples of N=5,000 participants and performing LCA using 1 through 5 clusters. We used the Bayesian Information Criterion and the elbow method to determine the optimal number of clusters. Based on these analyses, we determined that a 3-cluster solution was as the optimal in all 10 iterations. We further re-ran LCA in the complete dataset using only this solution, setting the maximum number of iterations through each estimation algorithm (maxiter) as 1,000 and the number of times to estimate the model with different class-conditional response probabilities (nrep) as 25, with default parameters otherwise. We used the poLCA package in R[30]. Cross-sectional associations between social risk factor clusters and OSA were assessed using chi-squared test and unadjusted and adjusted logistic regression. Covariates included age, sex, language, race, marital status, ethnicity, and urban/rural status.

We proceeded to determine whether different social risk factor clusters would prioritize different clinical risk factors towards predicting MACE risk among a cohort of individuals with evidence of OSA. For this analysis, we included only participants with evidence of OSA, at least 5 years of

follow-up data, to allow for ascertainment of MACE incidence. Phecode feature sets were used as predictors of 5-year incidence of MACE (binary outcome) using STREAMLINE, an end-to-end rigorous and interpretable auto-ML pipeline (https://github.com/UrbsLab/STREAMLINE)[31,32], which has been implemented in a SageMaker instance of the AIM-AHEAD Service Workbench. Data were split into training/testing (90%) and validation (10%), maintaining proportions for both the outcome and social risk factor clusters. For each cluster, we optimized four different ML methods (logistic regression [LR], random forest [RF], Light Gradient Boost Machine [LightGBM], and Extreme Gradient Boosting [XGB]), as well as evaluated models with area under the receiver operating characteristics curve (ROC-AUC) and area under the precision-recall curve (PRC-AUC) using a 3-fold cross-validation design. Feature importance scores were determined, along with social risk factor cluster-specific final models for independent validation. The top performing features in each subgroup were then selected and compared across clusters. Analyses were conducted using R (v 4.1.3) and Python (v 3.10.8).

## 3. Results

### 3.1. *Sample characterization*

In our initial analysis focused on assessing the association between social risk factors and prevalence of sleep disorders, our primary cohort consisted of 1,476,358 adults with encounters in community health centers across the U.S. **Figure 1** represent the study flowchart.

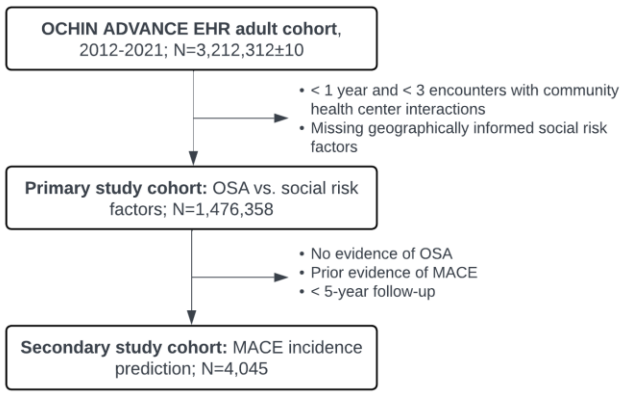


**Figure 1.** Study flowchart representing sample sizes for each included study cohort.

Among those, 63.2% were female, 69.9% spoke English as the primary language, 67.4% were White, 20.9% were Black, 5.1% were Asian, 67.8% had a current FPL <100%, 3.3% reported being homeless and 16.5% lived in rural areas. These characteristics highlight the sociodemographic diversity of the included cohort. **Table 1** provides descriptive statistics of the overall sample, as well as by evidence of OSA status. Individuals with evidence of an OSA diagnosis represented 2.3% of the included cohort (N=33,064), and univariate analyses suggest they were older, more likely to be males and with male gender identity, more likely to speak English as primary language, more likely to be White, less likely among those who were single, less likely among those with current FPL

<100%, less likely to be Hispanic or Latino, less likely among those reporting homelessness, more likely among those reporting heterosexual orientation, and more likely among those living in rural areas. Geographically informed social risk factors mostly differ between those with and without evidence of OSA, suggesting that those with a diagnosis are more likely to live in areas with lower social risk (**Table 1**).

**Table 1.** Sample characteristics, overall and stratified by evidence of obstructive sleep apnea (OSA).

| Variable | Category | Overall (N=1,476,358) | Evidence of OSA | | p[a] |
|---|---|---|---|---|---|
| | | | No (N=1,443,294) | Yes (N=33,064) | |
| **Age**, years | | 44.3 (15.9) | 44.1 (15.9) | 52.1 (12.8) | <0.001 |
| **Sex** | Female | 932,903 (63.2) | 917,179 (63.6) | 15,724 (47.6) | <0.001 |
| | Male | 543,036 (36.8) | 525,703 (36.4) | 17,333 (52.4) | |
| **Primary language** | English | 1,031,528 (69.9) | 1,002,947 (69.5) | 28,581 (86.4) | <0.001 |
| | Spanish | 354,107 (24.0) | 350,674 (24.3) | 3433 (10.4) | |
| | Other | 90,723 (6.1) | 89,673 (6.2) | 1,050 (3.2) | |
| **Race** | White | 993,964 (67.4) | 969,574 (67.3) | 24,390 (73.9) | <0.001 |
| | American Indian or Alaska Native | 12,696 (0.9) | 12,335 (0.9) | 361 (1.1) | |
| | Asian | 75,552 (5.1) | 74,646 (5.2) | 906 (2.7) | |
| | Black or African American | 307,628 (20.9) | 301,653 (20.9) | 5,975 (18.1) | |
| | Multiple Race | 20,259 (1.4) | 19,774 (1.4) | 485 (1.5) | |
| | Native Hawaiian or Other Pacific Islander | 9,899 (0.7) | 9,622 (0.7) | 277 (0.8) | |
| | Refuse to answer | 54,396 (3.7) | 53,771 (3.7) | 625 (1.9) | |
| **Marital status** | Current Partnership | 363,381 (24.6) | 355,745 (24.6) | 7,636 (23.1) | <0.001 |
| | Divorced/Separated | 86,176 (5.8) | 83,422 (5.8) | 2,754 (8.3) | |
| | Single | 496,312 (33.6) | 488,703 (33.9) | 7,609 (23.0) | |
| | Unknown | 497,609 (33.7) | 483,544 (33.5) | 14,065 (42.5) | |
| | Widowed | 32,880 (2.2) | 31,880 (2.2) | 1,000 (3.0) | |
| **Current FPL** | 101-150 % | 217,701 (14.7) | 212,929 (14.8) | 4,772 (14.4) | <0.001 |
| | ≤100 % | 1,000,448 (67.8) | 979,666 (67.9) | 20,782 (62.9) | |
| | 151-200 % | 94,042 (6.4) | 91,871 (6.4) | 2,171 (6.6) | |
| | >200 % | 164,167 (11.1) | 158,828 (11.0) | 5,339 (16.1) | |
| **Ethnicity** | Not Hispanic or Latino | 912,928 (62.8) | 886,753 (62.4) | 26,175 (80.5) | <0.001 |
| | Hispanic or Latino | 540,075 (37.2) | 533,727 (37.6) | 6,348 (19.5) | |
| **Gender identity** | Female | 535,485 (36.3) | 523,697 (36.3) | 11,788 (35.7) | <0.001 |
| | Male | 316,468 (21.4) | 304,107 (21.1) | 12,361 (37.4) | |
| | Transgender, Gender Queer, Other | 19,768 (1.3) | 19,591 (1.4) | 177 (0.5) | |
| | Unknown | 604,637 (41.0) | 595,899 (41.3) | 8,738 (26.4) | |
| **Homelessness status** | No/Unknown | 1,427,117 (96.7) | 1,394,342 (96.6) | 32,775 (99.1) | <0.001 |
| | Yes | 49,241 (3.3) | 48,952 (3.4) | 289 (0.9) | |
| **Sexual orientation** | Heterosexual | 731,141 (49.5) | 710,476 (49.2) | 20,665 (62.5) | <0.001 |
| | Homosexual | 24,422 (1.7) | 23,584 (1.6) | 838 (2.5) | |
| | Bisexual | 15,222 (1.0) | 14,837 (1.0) | 385 (1.2) | |
| | Other | 6,006 (0.4) | 5,856 (0.4) | 150 (0.5) | |
| | Unknown | 699,567 (47.4) | 688,541 (47.7) | 11,026 (33.3) | |
| **Rural/urban status** | Urban | 1,233,127 (83.5) | 1,209,391 (83.8) | 23,736 (71.8) | <0.001 |
| | Rural | 243,231 (16.5) | 233,903 (16.2) | 9,328 (28.2) | |
| **Geographically informed indicators** | | | | | |
| Unemployment rate, % | | 7.33% (3.41) | 7.33% (3.42) | 7.14% (3.29) | <0.001 |
| Median household income, U.S. dollars | | $53,994 (20,242) | $53,968.92 (20,269) | $55,086 (19,020) | <0.001 |
| % of college graduates | | 26.91% (15.06) | 26.91% (15.07) | 26.86% (14.63) | 0.518 |
| Gini coefficient | | 0.45 (0.05) | 0.45 (0.05) | 0.45 (0.05) | <0.001 |
| % of population below FPL | | 18.89 (9.37) | 18.92 (9.38) | 17.68 (8.68) | <0.001 |

[a] Chi-squared tests or t-tests. Categorical variables are represented as N (%) and continuous variable as mean (SD).
Abbreviations: OSA, obstructive sleep apnea; FPL, Federal Poverty Level; SD: standard deviation.

### 3.2. *Clusters of social risk factors*

Results of LCA revealed three major social risk factor clusters: lowest (N=489,191; 35.7%), average (N= 642,973; 43.6%) and highest (N=335,194; 22.7%) social burden. **Table 2** describes the differences between each geographically informed social risk factor and the 3-cluster solution used to inform the names of each cluster. The highest social burden cluster had the greatest proportion of the highest quartiles of unemployment rates, Gini coefficient, and proportion of individuals living below poverty level, and the lowest quartiles of median household income and proportion of individuals that are college graduates.

**Table 2.** Association between social risk factor quartiles and identified social risk clusters.

| Social risk factor quartiles | Category | Lowest Social Burden (35.7%) | Average Social Burden (43.6%) | Highest Social Burden (22.7%) | p[a] |
|---|---|---|---|---|---|
| **Unemployment rate** | Q1 [<5.1%] | 299,777 (60.2) | 61,207 (9.5) | 8,503 (2.5) | <0.001 |
| | Q2 [5.1-6.6%] | 139,264 (28.0) | 202,799 (31.5) | 23,253 (6.9) | |
| | Q3 [6.6-8.9%] | 51,053 (10.2) | 267,651 (41.6) | 49,361 (14.7) | |
| | Q4 [≥8.9%] | 8,097 (1.6) | 111,316 (17.3) | 254,077 (75.8) | |
| **Median household income** | Q1 [<40.7k] | 424 (0.1) | 56,959 (8.9) | 313,600 (93.6) | <0.001 |
| | Q2 [40.7-50.0k] | 1,856 (0.4) | 343,123 (53.4) | 21,594 (6.4) | |
| | Q3 [50.0-63.8k] | 135,449 (27.2) | 230,266 (35.8) | <11 | |
| | Q4 [≥63.8k] | 360,462 (72.4) | 12,625 (2.0) | <11 | |
| **% of college graduates** | Q1 [<16.7%] | 17,317 (3.5) | 144,164 (22.4) | 205,989 (61.5) | <0.001 |
| | Q2 [16.7-23.1%] | 46,037 (9.2) | 247,449 (38.5) | 74,078 (22.1) | |
| | Q3 [23.1-33.5%] | 146,377 (29.4) | 179,683 (27.9) | 39,874 (11.9) | |
| | Q4 [≥33.5%] | 288,460 (57.9) | 71,677 (11.1) | 15,253 (4.6) | |
| **Gini coefficient** | Q1 [<0.42] | 204,591 (41.1) | 140,407 (21.8) | 20,896 (6.2) | <0.001 |
| | Q2 [0.42-0.45] | 104,607 (21.0) | 199,895 (31.1) | 63,231 (18.9) | |
| | Q3 [0.45-0.48] | 93,785 (18.8) | 169,999 (26.4) | 104,347 (31.1) | |
| | Q4 [≥0.48] | 95,208 (19.1) | 132,672 (20.6) | 146,720 (43.8) | |
| **% below poverty level** | Q1 [<12.1%] | 355,133 (71.3) | 1,4870 (2.3) | 1,067 (0.3) | <0.001 |
| | Q2 [12.1-17.6%] | 130,118 (26.1) | 236,982 (36.9) | <11 | |
| | Q3 [17.6-23.7%] | 4,322 (0.9) | 349,902 (54.4) | 14,214 (4.2) | |
| | Q4 [≥23.7%] | 8,618 (1.7) | 41,219 (6.4) | 319,913 (95.4) | |

[a] Chi-squared tests. Categorical variables are represented as N (%).

### 3.3. *Associations between social risk factors clusters and OSA*

We proceeded to determine the association between social risk factors clusters and evidence of OSA. Univariate analysis indicated that individuals with evidence of OSA were less likely to belong to the highest social burden cluster (16.8%) when compared to those without evidence of OSA (22.8%, p<0.001). On the other hand, those with evidence of OSA were more likely to belong to both the lowest and average social burden clusters when compared to those without evidence of OSA (34.9% vs. 33.7% and 48.3% vs. 43.4%, respectively, both p<0.001). Logistic regression adjusted for relevant confounders, including individual level social risk factors, indicated that individuals belonging to the lowest social burden cluster were less likely to have received a diagnosis of OSA (OR [95%CI] = 0.85 [0.82-0.88]) when compared to those belonging to the highest social burden cluster. On the other hand, individuals belonging to the average social risk burden were slightly more likely to have received a diagnosis of OSA (1.03 [1.01-1.06]) compared to those in the highest

social burden cluster. Results suggest important socio-environmental contributions to potential disparities in the diagnosis of OSA in community health centers.

### 3.4. *MACE prediction among individuals with OSA*

Next, we proceeded to understand the clinical factors contributing to increased CV risk among individuals with OSA in the included sample, without taking into consideration their social risk cluster. A cohort of 4,045 individuals with OSA, without prior evidence of MACE and with at least 5 years of follow-up since their first OSA diagnosis was included in this analysis. Among those, 327 (8.1%) individuals had evidence of a MACE within the 5-year follow-up.

Using a robust ML pipeline, we proceeded to create our training (90%) and testing (10%) sets, maintaining the proportions of incident MACE cases and social risk clusters. Our training dataset consisted of 3,641 individuals (294 [8.1%] cases) and our testing dataset consisted of 404 individuals (33 [8.2%] cases). We determined these training/testing splits to allow greater representation of the dataset during training, due to the limited sample size of the cohort.

First, we assessed the performance of clinical risk factors (represented as phecodes) to predict incident MACE in the training and testing datasets, regardless of social risk cluster membership, using four different ML methods (LR, RF, LightGBM, and XGB). **Figure 2** summarizes the prediction performances in terms of ROC-AUC and PRC-AUC across the four methods. While XGB demonstrated the best performance in the training dataset for both performance metrics (mean [SD across cross-validation] ROC-AUC = 0.67 [0.03]; PRC-AUC = 0.14 [0.02]), LR was the best performing method in the testing dataset (ROC-AUC = 0.70 [0.03]; PRC-AUC = 0.19 [0.02]).

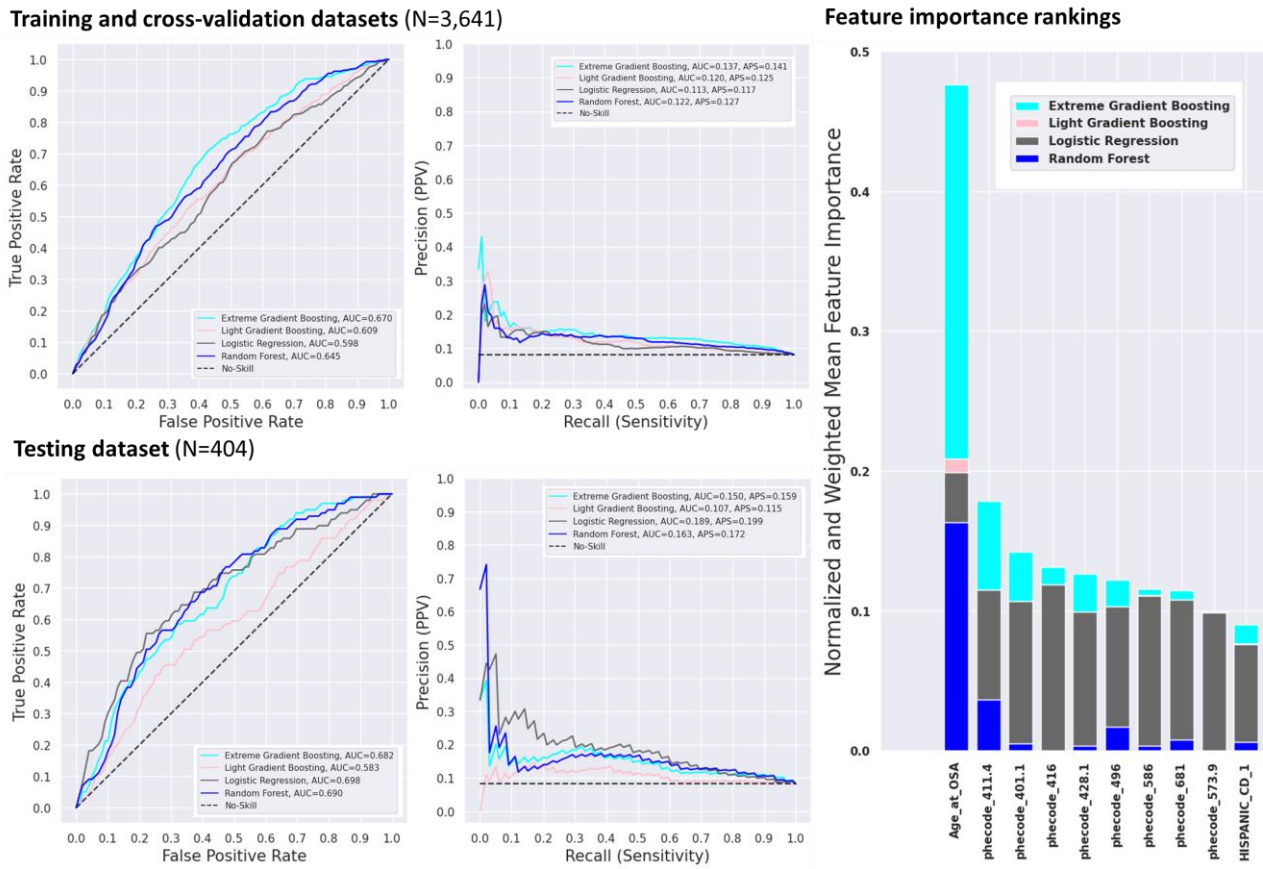


**Figure 2.** Summary of incident MACE prediction performance and feature importance (top 10 features).

Inspection of normalized and balanced accuracy-weighted feature importance plots (**Figure 2**) indicates that age was the most important predictor across all methods, except for LR. For this method, the most important feature was phecode 416 (cardiomegaly). Other relevant features listed among the top 10 included phecodes 411.4 (coronary atherosclerosis), 401.1 (essential hypertension), 428.1 (congestive heart failure), 496 (chronic airway obstruction), 586 (other disorders of the kidney and ureters), 681 (superficial cellulitis and abscess), 573.9 (abnormal serum enzyme levels) and ethnicity.

### 3.5. *MACE prediction after social risk factor cluster stratification*

Finally, we proceeded to explore how these models would perform within specific subgroups according to the assigned social risk factor clusters, and whether top clinical predictors would be similar or different across clusters. For this analysis we trained and evaluated ML models using the same methods described above, but within each social risk factor cluster. Training and testing dataset sample sizes for each cluster were as follows: lowest social burden cluster ($N_{train}$=1,136; $N_{test}$ = 126), average social burden cluster ($N_{train}$ =1,791; $N_{test}$ = 199), and highest social burden cluster ($N_{train}$ = 467; $N_{test}$ = 52).

**Table 3** summarizes the results of the predictive performance in the testing dataset from models trained and evaluated within each social risk factor cluster separately. According to the ROC-AUC, within participants assigned to the lowest social burden clusters, LR was the best performing method, while RF performed the best in both the average and highest social burden clusters. According to the PRC-AUC, within participants assigned to the lowest social burden clusters, LR was also the best performing method, while XGB performed the best in both the average and highest social burden clusters.

**Table 3.** Summary of prediction performance metrics in the testing datasets using models trained within social risk factor clusters.

| Method | Metric | Cluster | | |
|---|---|---|---|---|
| | | Lowest social burden | Average social burden | Highest social burden |
| **XGB** | ROC-AUC | 0.500 | 0.617 | 0.564 |
| | PRC-AUC | 0.117 | 0.133 | 0.189 |
| **LightGBM** | ROC-AUC | 0.616 | 0.606 | 0.504 |
| | PRC-AUC | 0.126 | 0.098 | 0.177 |
| **LR** | ROC-AUC | 0.689 | 0.631 | 0.522 |
| | PRC-AUC | 0.213 | 0.114 | 0.103 |
| **RF** | ROC-AUC | 0.634 | 0.634 | 0.628 |
| | PRC-AUC | 0.163 | 0.127 | 0.141 |

Abbreviations: XGB, Extreme Gradient Boosting; LightGBM, Light Gradient Boost Machine; LR, logistic regression; RF, random forest; ROC-AUC, area under the receiver operating characteristics curve; ROC-PRC, area under the precision-recall curve.

We then inspected differences in the normalized and balanced accuracy-weighted feature importance plots (**Figure 3**) across the models and social risk factor clusters to investigate whether clinical risk factors that predict incident MACE would be different depending on individuals' socio-economic exposures. Results suggest that while age at diagnosis of OSA was an important predictor across all social risk factor clusters, being listed among the top 10 features in all groups, there were important differences in the comorbidity profile linked to incident MACE within each group. For example, among those with lowest social burden, some more conventional CV comorbidities or risk factors were observed, such as essential hypertension (401.1), nonspecific chest pain (418) and both type 1 and 2 diabetes (250.1 and 250.2). However, among those with highest social burden, top predictors included symptoms such as malaise and fatigue (798), pain in joint (745), and dizziness and giddiness (light-headedness and vertigo, 386.9), in addition to a more metabolic comorbidity profile (244.4, hypothyroidism and 250.2, type 2 diabetes). Among those with average social burden, features included both conventional ones (401.1, essential hypertension and 416, cardiomegaly) as well as other infectious and parasitic diseases (136) and Lyme disease (130.1). Anxiety disorder (300.1) was also observed as an important predictor among those with lowest and average social burden.

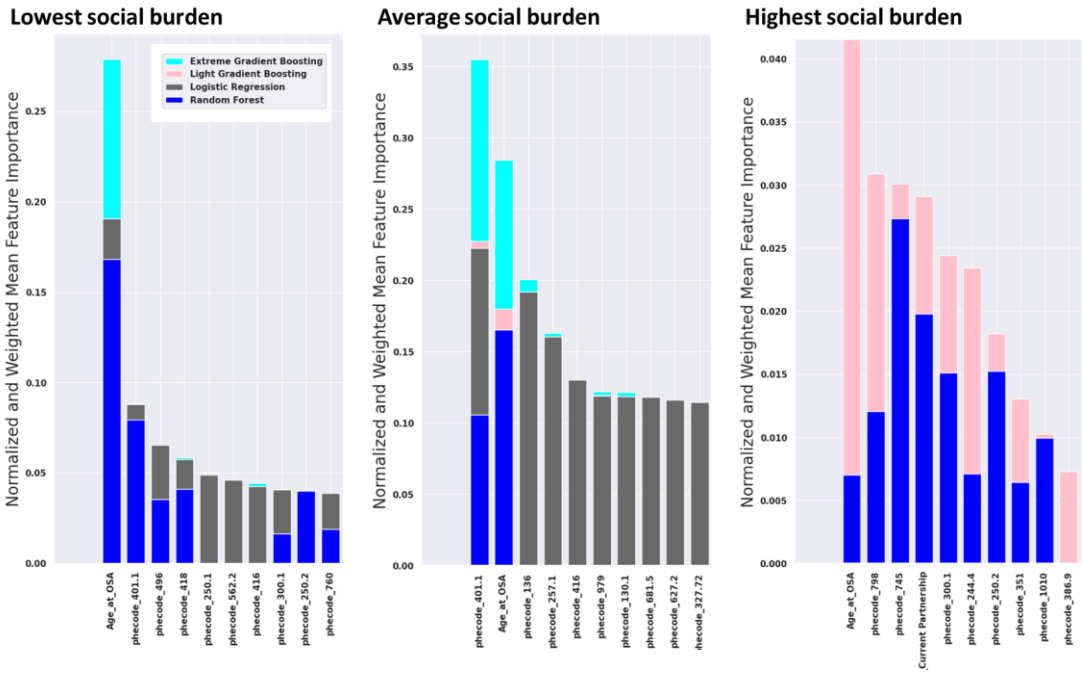


**Figure 3.** Top feature important comparison across models evaluated within different social risk factor clusters.

## 4. Discussion

Our main findings highlight important social disparities related to the identification and diagnosis of OSA in community health centers, as well as important differences in clinical factors that contributed to the prediction of incident CV diseases among participants with a diagnosis of OSA. We applied an innovative approach to identify social risk factor clusters derived from relevant geographically informed social indicators estimated from national surveys. We identified three clusters (lowest, average, and highest social burden), consistent with observed individual-level sociodemographic characteristics. Individuals belonging to the highest social burden cluster were less likely to have received a diagnosis of OSA, even after adjusting for relevant confounders such as sex, race, and ethnicity – factors that have been consistently demonstrated to affect health disparities within sleep disorders[1-8]. Our study also demonstrated that a LR-based incident MACE prediction model trained on hundreds of clinical features (i.e., phecodes) had a reasonable, yet not optimal performance in testing sets. Nevertheless, performance varied across subgroups defined by social risk factor clusters, as well as the top features contributing to those predictions, suggesting different pathways towards CV risk depending on socio-environmental exposure.

The study provides novel insights about the clinical prevalence and recognition of OSA within community health centers in a diverse population at greater social burden. Our dataset was composed of a large proportion of underrepresented minorities according to sociodemographic characteristics, including race, ethnicity, gender and sexual identity. More importantly, 67.8% of the cohort were below the Federal poverty level. In this context, the observed clinical prevalence of OSA (2.3%) is lower than other clinical cohorts defined using EHR-based methods, such as within the National COVID Cohort Collaborative (3.9%), comprised of a sample of individuals that have

been tested positively for SARS-CoV-2 through encounters within academic health systems[33]. The prevalence is even lower than the expected population prevalence of OSA, estimated to affect nearly 1 billion people worldwide[12]. It is well-established that OSA is underdiagnosed[34,35], and our analysis in community health centers identified even further differences. Clustering of geographically informed social indicators revealed that individuals at greater social burden (i.e., highest social burden cluster) were significantly less likely to have received a diagnosis of OSA, even after adjusting for other established individual-level social risk factors. When assessing individual-level sociodemographic characteristics, those receiving a diagnosis of OSA were more likely to speak English as primary language, more likely to be White, less likely to be among those with current FPL <100%, less likely to be Hispanic or Latino, less likely to report homelessness, although more likely to live in rural areas. These findings suggest important socio-environmental contributions to potential disparities in the diagnosis of OSA in community health centers and that underrepresented minorities may not be receiving adequate sleep care. Thus, screening of sleep disorders particularly within this subgroup at greater risk is necessary. While it might seem impractical to offer screening and treatment of chronic sleep disorders such as OSA in community health centers, preventing high risk individuals from obtaining access to quality sleep health care might exacerbate disparities related to metabolic, neurological, and psychiatric conditions, all of which have been associated with OSA[36].

In this context, CV diseases are particularly relevant due to the worsening of CV disparities over several decades, despite efforts of addressing health needs of vulnerable populations[37]. Due to the major epidemiological and experimental evidence supporting the role of OSA towards increasing CV risk[13-16], understanding and addressing these needs are of high importance. Towards this goal, we assessed whether an incident MACE prediction model trained on a broad range of clinical features within individuals with OSA had adequate performance and could be used to prioritize clinical profiles based on most relevant features. Despite our best model, a LR with a ROC-AUC of 0.70 and a PRC-AUC of 0.19, not being necessarily optimal for deployment, it helped us identify important features contributing to the prediction, many of them with established associations with OSA. For example, in our overall analysis, top features included cardiomegaly, atherosclerosis, essential hypertension, congestive heart failure, chronic airway obstruction, disorders of the kidney and ureters, and abnormal serum enzyme levels. Many of these features are established CV risk factors, supporting internal validity of our approach. More importantly, therapies focused on mitigating the effects of OSA have been demonstrated to improve some of these risk factors[38].

When assessing the prediction performance of models across strata of social risk factor clusters, we continued to identify similar, although slightly lower performance across groups with testing ROC-AUC ranging from 0.63 to 0.69 and PRC-AUC ranging from 0.13 to 0.21 for the best models. This is likely explained by the smaller sample size used for training in the stratified analyses, preventing models from learning relationships between clinical features and outcome. Some key clinical factors contributing to these predictions are observed across all social risk factor clusters, such as age at OSA diagnosis, cardiometabolic conditions (e.g., type 2 diabetes, hypertension), and anxiety disorders. However, among those with highest social burden, top predictors included symptom-related factors such as malaise and fatigue, pain in joint, and light-headedness and vertigo, while among those with average social burden, features included infectious and parasitic diseases. These presentations might reflect primary reasons or exposure to different healthcare specialists. In this context, the study provided a systematic data driven approach to identify these factors, where

future studies could further explore, under a more hypothesis-driven methodology whether these conditions could be suggestive of higher CV risk within vulnerable populations.

Our study has several strengths, such as providing an analysis in a large, racial, ethnic, and socioeconomically diverse clinical cohort of individuals observed in community health centers, a target population often neglected from epidemiological and experimental studies. In addition, we use a robust ML pipeline comparing, in a systematic way, different sets of ML methods and features towards understanding clinical factors of incident CV diseases. However, our study also present important limitations that should be considered when interpreting the findings. Access to OSA therapies, such as continuous positive airway pressure or mandibular advancement devices are likely not offered by this care modality and therefore not necessarily recorded in the ADVANCE EHR data warehouse, thus they could not be considered as confounders. More granular information about severity of OSA based on the apnea-hypopnea index or other metrics was not available, as it required parsing of clinical sleep study reports. Similarly, phecodes are not necessarily always precise, granular measures of diagnoses and may lack sensitivity and specificity of validated computable phenotypes. However, as part of a data-driven EHR-wide analysis, they may offer an initial set of hypotheses that could be assessed with more robust phenotypes in future investigations. Despite our observed signals, incidence rates of MACE were relatively low, possibly due to the relative short, 5-year follow-up time, resulting in a very imbalanced classification problem. However, longer follow-up windows would substantially reduce sample size and was not a feasible alternative.

In conclusion, this study leveraged heterogeneous EHR data from community health centers in the United States and described sociodemographic and geographically informed social disparities as they relate to diagnosis of OSA. Prediction models of incident MACE among individuals experiencing OSA also disparities in across clinical predictors of CV diseases. Thus, tailored interventions geared toward minimizing these disparities are warranted.

## 5. Acknowledgments

The research reported in this work was powered by PCORnet®. PCORnet has been developed with funding from the Patient-Centered Outcomes Research Institute® (PCORI®) and conducted with the Accelerating Data Value Across a National Community Health Center Network (ADVANCE) Clinical Research Network (CRN). ADVANCE is a Clinical Research Network in PCORnet® led by OCHIN in partnership with Health Choice Network, Fenway Health, University of Washington, and Oregon Health & Science University. ADVANCE's participation in PCORnet® is funded through the PCORI Award RI-OCHIN-01-MC.

## References


1.Jackson CL, Walker JR, Brown MK, Das R, Jones NL. A workshop report on the causes and consequences of sleep health disparities. *Sleep*. Aug 12 2020;43(8)doi:10.1093/sleep/zsaa037
2.Carnethon MR, De Chavez PJ, Zee PC, et al. Disparities in sleep characteristics by race/ethnicity in a population-based sample: Chicago Area Sleep Study. *Sleep Med*. Feb 2016;18:50-5. doi:10.1016/j.sleep.2015.07.005
3.Jackson CL, Patel SR, Jackson WB, 2nd, Lutsey PL, Redline S. Agreement between self-reported and objectively measured sleep duration among white, black, Hispanic, and Chinese adults in the United States: Multi-Ethnic Study of Atherosclerosis. *Sleep*. Jun 1 2018;41(6)doi:10.1093/sleep/zsy057
4.Johnson DA, Lisabeth L, Hickson D, et al. The Social Patterning of Sleep in African Americans: Associations of Socioeconomic Position and Neighborhood Characteristics with Sleep in the Jackson Heart Study. *Sleep*. Sep 1 2016;39(9):1749-59. doi:10.5665/sleep.6106
5.Kaufmann CN, Mojtabai R, Hock RS, et al. Racial/Ethnic Differences in Insomnia Trajectories Among U.S. Older Adults. *Am J Geriatr Psychiatry*. Jul 2016;24(7):575-84. doi:10.1016/j.jagp.2016.02.049
6.Liu Y, Wheaton AG, Chapman DP, Cunningham TJ, Lu H, Croft JB. Prevalence of Healthy Sleep Duration among Adults--United States, 2014. *MMWR Morb Mortal Wkly Rep*. Feb 19 2016;65(6):137-41. doi:10.15585/mmwr.mm6506a1
7.Petrov ME, Lichstein KL. Differences in sleep between black and white adults: an update and future directions. *Sleep Med*. Feb 2016;18:74-81. doi:10.1016/j.sleep.2015.01.011
8.Roane BM, Johnson L, Edwards M, Hall J, Al-Farra S, O'Bryant SE. The link between sleep disturbance and depression among Mexican Americans: a Project FRONTIER study. *J Clin Sleep Med*. Apr 15 2014;10(4):427-31. doi:10.5664/jcsm.3622
9.Dudley KA, Patel SR. Disparities and genetic risk factors in obstructive sleep apnea. *Sleep Med*. Feb 2016;18:96-102. doi:10.1016/j.sleep.2015.01.015
10.Seixas AA, Trinh-Shevrin C, Ravenell J, Ogedegbe G, Zizi F, Jean-Louis G. Culturally tailored, peer-based sleep health education and social support to increase obstructive sleep apnea assessment and treatment adherence among a community sample of blacks: study protocol for a randomized controlled trial. *Trials*. Sep 24 2018;19(1):519. doi:10.1186/s13063-018-2835-9
11.Jean-Louis G, Newsome V, Williams NJ, Zizi F, Ravenell J, Ogedegbe G. Tailored Behavioral Intervention Among Blacks With Metabolic Syndrome and Sleep Apnea: Results of the MetSO Trial. *Sleep*. Jan 1 2017;40(1)doi:10.1093/sleep/zsw008
12.Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med*. Aug 2019;7(8):687-698. doi:10.1016/S2213-2600(19)30198-5
13.Drager LF, Togeiro SM, Polotsky VY, Lorenzi-Filho G. Obstructive Sleep Apnea. *Journal of the American College of Cardiology*. 2013;62(7):569-576. doi:10.1016/j.jacc.2013.05.045
14.Mazzotti DR, Keenan BT, Lim DC, Gottlieb DJ, Kim J, Pack AI. Symptom Subtypes of Obstructive Sleep Apnea Predict Incidence of Cardiovascular Outcomes. *Am J Respir Crit Care Med*. Aug 15 2019;200(4):493-506. doi:10.1164/rccm.201808-1509OC
15.Azarbarzin A, Sands SA, Stone KL, et al. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the Osteoporotic Fractures in Men Study and the Sleep Heart Health Study. *European Heart Journal*. 2019-04-07 2019;40(14):1149-1157. doi:10.1093/eurheartj/ehy624

16.Azarbarzin A, Sands SA, Younes M, et al. The Sleep Apnea–Specific Pulse-Rate Response Predicts Cardiovascular Morbidity and Mortality. *American Journal of Respiratory and Critical Care Medicine*. 2021;203(12):1546-1555. doi:10.1164/rccm.202010-3900OC
17.DeVoe JE, Gold R, Cottrell E, et al. The ADVANCE network: accelerating data value across a national community health center network. *J Am Med Inform Assoc*. Jul-Aug 2014;21(4):591-5. doi:10.1136/amiajnl-2014-002744
18.Hughes LS, Phillips RL, DeVoe JE, Bazemore AW. Community Vital Signs: Taking the Pulse of the Community While Caring for Patients. *The Journal of the American Board of Family Medicine*. 2016;29(3):419-422. doi:10.3122/jabfm.2016.03.150172
19.Tan AX, Hinman JA, Abdel Magid HS, Nelson LM, Odden MC. Association Between Income Inequality and County-Level COVID-19 Cases and Deaths in the US. *JAMA Network Open*. 2021;4(5)doi:10.1001/jamanetworkopen.2021.8799
20.De Maio FG. Income inequality measures. *Journal of Epidemiology & Community Health*. 2007;61(10):849-852. doi:10.1136/jech.2006.052969
21.Keenan BT, Kirchner HL, Veatch OJ, et al. Multisite validation of a simple electronic health record algorithm for identifying diagnosed obstructive sleep apnea. *J Clin Sleep Med*. Feb 15 2020;16(2):175-183. doi:10.5664/jcsm.8160
22.Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annual Review of Biomedical Data Science*. 2021;4(1):1-19. doi:10.1146/annurev-biodatasci-122320-112352
23.Singer DE, Chang Y, Borowsky LH, et al. A New Risk Scheme to Predict Ischemic Stroke and Other Thromboembolism in Atrial Fibrillation: The ATRIA Study Stroke Risk Score. *Journal of the American Heart Association*. 2013;2(3)doi:10.1161/jaha.113.000250
24.Go AS, Hylek EM, Chang Y, et al. Anticoagulation Therapy for Stroke Prevention in Atrial Fibrillation. *Jama*. 2003;290(20)doi:10.1001/jama.290.20.2685
25.Sidney S, Sorel M, Quesenberry CP, DeLuise C, Lanes S, Eisner MD. COPD and Incident Cardiovascular Disease Hospitalizations and Mortality: Kaiser Permanente Medical Care Program. *Chest*. 2005;128(4):2068-2075. doi:10.1378/chest.128.4.2068
26.Go AS, Yang J, Ackerson LM, et al. Hemoglobin Level, Chronic Kidney Disease, and the Risks of Death and Hospitalization in Adults With Chronic Heart Failure. *Circulation*. 2006;113(23):2713-2723. doi:10.1161/circulationaha.105.577577
27.Gurwitz JH, Magid DJ, Smith DH, et al. Contemporary Prevalence and Correlates of Incident Heart Failure with Preserved Ejection Fraction. *The American Journal of Medicine*. 2013;126(5):393-400. doi:10.1016/j.amjmed.2012.10.022
28.McKee PA, Castelli WP, McNamara PM, Kannel WB. The Natural History of Congestive Heart Failure: The Framingham Study. *New England Journal of Medicine*. 1971;285(26):1441-1446. doi:10.1056/nejm197112232852601
29.Chen W, Yao J, Liang Z, et al. Temporal Trends in Mortality Rates among Kaiser Permanente Southern California Health Plan Enrollees, 2001-2016. *Perm J*. 2019;23doi:10.7812/TPP/18-213
30.Linzer DA, Lewis JB. **poLCA** : An *R* Package for Polytomous Variable Latent Class Analysis. *Journal of Statistical Software*. 2011 2011;42(10)doi:10.18637/jss.v042.i10
31.Urbanowicz R, Zhang R, Cui Y, Suri P. STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison. *Genetic Programming Theory and Practice XIX*. 2023:201-231:chap Chapter 9. *Genetic and Evolutionary Computation*.

32.Urbanowicz RJ, Bandhey H, Keenan BT, et al. STREAMLINE: An Automated Machine Learning Pipeline for Biomedicine Applied to Examine the Utility of Photography-Based Phenotypes for OSA Prediction Across International Sleep Centers. 2023:arXiv:2312.05461. doi:10.48550/arXiv.2312.05461 Accessed December 01, 2023. https://ui.adsabs.harvard.edu/abs/2023arXiv231205461U
33.L Mandel H, Colleen G, Abedian S, et al. Risk of post-acute sequelae of SARS-CoV-2 infection associated with pre-coronavirus disease obstructive sleep apnea diagnoses: an electronic health record-based analysis from the RECOVER initiative. *Sleep*. 2023;46(9)doi:10.1093/sleep/zsad126
34.Kapur V, Strohl KP, Redline S, Iber C, O?Connor G, Nieto J. Underdiagnosis of Sleep Apnea Syndrome in U.S. Communities. *Sleep and Breathing*. 2002;6(2):49-54. doi:10.1055/s-2002-32318
35.Finkel KJ, Searleman AC, Tymkew H, et al. Prevalence of undiagnosed obstructive sleep apnea among adult surgical patients in an academic medical center. *Sleep Medicine*. 2009;10(7):753-758. doi:10.1016/j.sleep.2008.08.007
36.Gleeson M, McNicholas WT. Bidirectional relationships of comorbidity with obstructive sleep apnoea. *European Respiratory Review*. 2022;31(164)doi:10.1183/16000617.0256-2021
37.Walton-Moss B, Samuel L, Nguyen TH, Commodore-Mensah Y, Hayat MJ, Szanton SL. Community-Based Cardiovascular Health Interventions in Vulnerable Populations. *Journal of Cardiovascular Nursing*. 2014;29(4):293-307. doi:10.1097/JCN.0b013e31828e2995
38.Sircu V, Colesnic S-I, Covantsev S, et al. The Burden of Comorbidities in Obstructive Sleep Apnea and the Pathophysiologic Mechanisms and Effects of CPAP. *Clocks & Sleep*. 2023;5(2):333-349. doi:10.3390/clockssleep5020025